SciMedCentral

Annals of Biometrics & Biostatistics

Review Article

Statistical Models to Analyze Genotoxicological Experiments with the Comet Assay Data

Pablo Emilio Verde* and Anika Rottmann

Coordination Center for Clinical Trials, University of Düsseldorf, Germany

***Corresponding author**

Pablo E. Verde, Coordination Center for Clinical Trials, University of Düsseldorf Moorenstr, 540225 Düsseldorf, Germany, Tel: 492118104129; Email: pabloemilio.verde@hhu.de

Submitted: 06 November 2015

Accepted: 23 December 2015

Published: 26 December 2015

Keywords

- Comet assay
- Genotoxicity
- Random effects
- Hierarchical models
- Bayesian analysis

Abstract

The comet assay is a simple, rapid and sensitive tool for direct visualization of DNA damage in individual cells. This assay is commonly applied in regulatory genotoxicological studies and environmental monitoring. It is well known that data generated by comet assays are difficult to analyze. Typically the distributions of outcomes are asymmetric and they don't follow standard parametric distributions. To complicate matters, some data may present an excess of zeros and observations are nested within two to three levels in each experimental unit. During the last years a series of innovative statistical models have been used to analyze comet assay data. In this review we bring together these models by highlighting their advantages and disadvantages.

INTRODUCTION

The comet assay is a simple, rapid and sensitive tool for direct visualization of DNA damage in individual cells [1]. This assay is commonly applied in regulatory genotoxicological studies and environmental monitoring.

The basis of the comet assay involves embedding a suspension of single cells in low melting point agarose on a microscope slide. The cells are lysed and electrophoresis is carried out, when subjected to an electric field, the DNA migrates out of the cell, in the direction of the anode. The image obtained after electrophoresis looks like a "comet" with a clear head, corresponding to intact DNA, and a tail, consisting of damaged pieces of DNA; hence the name "Comet Assay".

The extent of DNA strand breakage can be analyzed by using image analysis systems that provide detailed information, such as comet tail length, % of DNA in the comet tail and tail moment (product of tail length and % of DNA in the comet tail). It is well known that these data are difficult to analyze. Typically, the distributions of these outcomes are asymmetric and they don't follow standard parametric distributions, even after transformation of the data as pointed out by Lovell *et al.,* [3]. To complicate matters, some data may present an excess of zeros, which may lead to bias in the analysis as described by Bright *et al.,* [4].

A typical experimental design of comet data includes two (or more) different treatments or dose groups. Experimental units (e.g. animals) are randomly assigned to one of the treatments. Samples of cells of each unit (typically two or three samples of 50 cells) are placed on slides to be analyzed. This design results in several different sources of variation: between experimental units, between slides within units and between cells within slides. In practice, the variability between slides is usually ignored, which results in two level nested data with cells within experimental units and units within treatment groups.

A detailed review on statistical issues concerning comet assay data is presented by Lovell and Omori [2]. However, during the last years a series of innovative statistical models have been used to analyze comet assay data. In this review we present these models by highlighting their advantages and disadvantages.

Strategies for analyzing comet data

There are two main strategies to analyze comet data: One is to analyze the data at the level of the cell and build a mixed effect model with fixed effect of treatment and random effects of animal and slide. This approach accounts for variability between and within animals and between slides within animals. An alternative is to take only animals and not slides as random effects in the model. The main problem of using mixed effects models is that the outcome variable is usually not Normal and with an excess of zeros, that may lead to difficulties in the analysis of the data.

An alternative approach is to aggregate the cell data at the level of the slide or at the level of the animals. Then the model only contains fixed effects and a simple statistical procedure (e.g. ANOVA) is used to make group comparisons.

Statistical procedures for aggregate comet assay data

Simple approaches to analyze comet data have been based on summaries at the level of the animal or experimental unit. In

their recommendations, the Comet Assay Interest Group (www.cometassay.com) has included the use of the median as a central parameter. The main reason is that the distribution of the data is asymmetric and does not follow a Normal distribution. For example, if 50 cells are analyzed per animal, then the median of these 50 values constitutes the summary of the animal outcome. This method ignores the possible variability between cells and between slides within animals.

Once the information within the animal is summarized, then a simple statistical procedure, like ANOVA, is used to compare the different groups. Duez *et al.,* [5] have described some usually employed statistical methods by using the log transformation of the comet measurement (e.g. tail length) or the data without transformation. Furthermore, for tail moment data, Wiklund and Agurell [6] recommend the use of the log transformation based on simulation experiments and power evaluations of statistical techniques.

In 2010, statisticians in the Pharmaceutical Industry (PSI) Toxicology Special Interest Group met to discuss the design and analysis of the comet assay. Bright *et al.,* [4] summarized the resulting recommendations. These include: 1) display the aggregate data at the level of the slide and animal; 2) use the log transformation of the comet outcome; 3) summarize observations at the level of the animal with the mean across replicated slides of the median of the log transformation of the tail intensity; 4) in case of having one third or more observations equal to zero, using two outcome summaries per animal is recommended; a) the proportion of non-zero % tail intensities and b) median of the log tail intensities excluding observations equal to zero.

Once the aggregated summary per animal is calculated, Bright *et al.,* [4] recommend using a simple statistical procedure. However, when several observations are equal to zero and two summaries are used per animal, they don't make further recommendations.

Statistical modeling approaches for individual cell data

As mentioned previously, data such as tail moment present skewed distributions, so a number of statistical distributions have been proposed for modeling comet data. Dehon *et al.,* [7] proposed to use the sum of two Gaussian distributions and Ejchart and Sadlej-Sosnowska [8] the Weibull distribution. However, these works were not able to model the clustering structure of the data.

Verde *et al.,* [9] were the first to parametrically model the skewed characteristic of comet data with the hierarchical structure of the data. In this work the authors modeled tail moment data using a two-level hierarchical model. In the first level of the model, observations are assumed to follow a distribution that belongs to the family of the accelerated life models. This family includes as special cases the following distributions: Weibull, exponential, logistic, normal, log-normal and log-logistic. In the second stage of the model a random effect, which follows a Gamma distribution, is used to account for the cluster structure of the data. The authors recommend choosing the distribution that fits best to the data.

Treatment effects were assessed by the construction of a "probability over damage" (POD) graph, which visualized the degree of damage produced by the treatment. POD is the probability that a tail moment (tm) measurement is greater than a given value "k", $Pr(tm > k)$. The horizontal axis of the graph corresponds to observed values of tail moment and the vertical axis to POD. A POD curve can be constructed for each individual in order to visualize individual variability, and also for each experimental group. A group with higher POD corresponds to larger damaged values. POD curves can also be used for exploratory analysis of the comet data. This plot corresponds to the complementary cumulative probability function and was an idea borrowed from survival analysis of time to event data. The authors provided examples of how to use the statistical software R to fit these models.

Li *et al.,* [10] proposed to use a hierarchical zero-inflated log-normal model. In this approach, the authors modeled a three-level hierarchical data from a longitudinal study, where cells are nested within subjects and subjects are measured on different occasions. By using a combination of two regression equations, the model simultaneously accounts for comet outcomes equal to zero and outcomes greater than zero. Two random components are used to explain the variability between subjects and visits. The authors propose to use an EM algorithm for parameter estimation and a bootstrapping technique for assessing parameters variability. They don't provide any software to apply their model.

Efendi and Molenberghs [11] proposed a multilevel model which is motivated by the work of Molenberghs *et al.,* [12]. In this approach, the authors assumed that the tail intensity percentage of the comet can be modeled with a Weibull distribution conditionally on two normally distributed random effects. These random effects are used to model two levels of clustering, slides within animals and cells within slides. The authors analyze three different estimation strategies: full likelihood, pseudo-likelihood and Bayesian estimation. The authors show that all three strategies work well and point out that while the pairwise likelihood estimation on the one hand needs more computation time than the others, it is, on the other hand, less sensitive to starting values. In order to assess joined dose effects in different comet assay outcomes, Efendi *et al.,* [13] extended the univariate model for the tail intensity percentage to a bivariate model for the tail length and tail intensity. The authors assume that there are two different sets of random effects: The first set is used to take into account the correlation between outcomes. The second set of random effects is added in order to model over dispersion.

One early work in modeling the complexities of the comet assay data was by Dunson *et al.,* [14]. They combined a median regression model with a hierarchical latent variable, which is used to handle the cluster structure of the data. In addition, they modeled several comet assay summaries simultaneously in a multivariate way. Latent response models typically assume that residual densities are Gaussian distributed; in this work the authors relaxed this assumption, hence their approach can be classified as a semi-parametric Bayesian modeling. Dunson [15] used the same data to illustrate an innovative full non-parametric Bayesian modeling based on a Dirichlet process for the latent structure of the data. Rodriguez [16] presented a Bayesian

hierarchical density regression model to analyze olive tail moment data. They used a finite mixture of Normal distributions, where they place regression equations to the mixture probabilities. The regression equations specified fixed effects corresponding to the treatment groups and random effects for the experimental units (e.g. animals, cell lines, etc.). This model can be considered as an approximation of the full non-parametric Bayesian approach proposed by Dunson [15]. The idea of placing a regression model on the mixture probabilities is borrowed from the latent class modeling literature. In these complex Bayesian hierarchical density approaches, Markov Chain Monte Carlo (MCMC) techniques are used to calculate posteriors of every model parameter. They used Bayes factors for model comparison, and for model fitness they compared predictive distribution for each individual against its corresponding empirical distribution.

DISCUSSION AND CONCLUSION

One important advantage of using aggregated data at the level of the animal is that it is easy to perform a simple statistical analysis and no specialized software is needed. In addition, the visualization of aggregated data is easy to interpret, because we are just comparing means and medians. The main disadvantage in aggregating data is the loss of information, for example we reduce 150 (e.g. three slides and 50 cells per slide) observations per animal to a single value. This means that we ignore the variability between slides within animals and cells within slides. Assuming that we can ignore the variability within animals is an extreme assumption. However, in a recent work, Hansen *et al.,* [17] used aggregated data at the level of the animal, but considered animals nested within doses as random effect, in addition the authors provided power calculations and R scripts. Using aggregated data with a mixed-effect model could be a practical solution to analyze these types of data.

Models based on parametric family distributions have the advantage to naturally extend the mixed effects models, which are very well known in statistical literature. Computations are also implemented in statistical software, but for some models [11,13] extra programing is needed. Interpretation of results is also simple. Some disadvantages of these approaches are that the comet data could present more complex features that are not modeled correctly, and the difficulty of handling multivariate outcomes.

For individual cell data, the Bayesian hierarchical density approach is in our opinion very interesting and could be the way to handle all the complexities of the comet data. In addition, they can easily be extended to multivariate data when we are interested in modeling different comet outcomes simultaneously. However, up to now only specialists can implement and use these models. Software which implements these techniques for non-specialists would be needed in order to use these models in practice.

We may ask which statistical methods are used in practice to analyze comet data. A recent review from the ComNet Project (Collins *et al.,* [18]) showed that very simple statistical techniques, which are not necessarily correctly applied, are commonly used in the majority of applications. In addition, results are presented by using p-values instead of estimates of treatment effects and their confidence intervals. This final remark is a warning that complex statistical models that have been developed during the last ten years should be implemented in user-friendly statistical software in order to enable practitioners to use these powerful statistical methods.

ACKNOWLEDGEMENTS

Pablo E. Verde is grateful to José Montserat for introducing him to the problem of analyzing comet assay data. This work was partially supported by the German Research Foundation projects DFG VE 896 /1-1.

REFERENCES

1. Ostling O, Johanson KJ. Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells. Biochem Biophys Res Commun. 1984; 123: 291-298.
2. Lovell DP, Omori T. Statistical issues in the use of the comet assay. Mutagenesis. 2008; 23: 171-182.
3. Lovell DP, Thomas G, Dubow R. Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. Teratogenesis, carcinogenesis, andmutagenesis. 1999; 19: 109-119.
4. Bright J, Aylott M, Bate S, Geys H, Jarvis P, Saul J, Vonk R. Recommendations on the statistical analysis of the Comet assay. Pharm Stat. 2011; 10: 485-493.
5. Duez P, Dehon G, Kumps A, Dubois J. Statistics of the Comet assay: a key to discriminate between genotoxic effects. Mutagenesis. 2003; 18: 159-166.
6. Wiklund SJ, Agurell E. Aspects of design and statistical analysis in the Comet assay. Mutagenesis. 2003; 18: 167-175.
7. Dehon G, Catoire L, Duez P, Bogaerts P, Dubois J. Validation of an automatic comet assay analysis system integrating the curve fitting of combined comet intensity profiles. Mutat Res. 2008; 650: 87-95.
8. Ejchart A, Sadlej-Sosnowska N. Statistical evaluation and comparison of comet assay results. Mutat Res. 2003; 534: 85-92.
9. Verde PE, Geracitano LA, Amado LL, Rosa CE, Bianchini A, Monserrat JM. Application of public-domain statistical analysis software for evaluation and comparison of comet assay data. Mutation Research/ Genetic Toxicology and Environmental Mutagenesis. 2006; 604: 71-82.
10. Ning Li, Elashoff DA, Robbins WA, Lin Xun. A hierarchical zero-inflated log-normal model for skewed responses. Stat Methods Med Res. 2011; 20: 175-189.
11. Efendi A, Molenberghs G. A multilevel model for hierarchical, repeated, and overdispersed time-to-event outcomes and its estimation strategies. J Biopharm Stat. 2013; 23: 1420-1434.
12. Ghebretinsae AH, Faes C, Molenberghs G, Geys H, Van der Leede BJ. Joint modeling of hierarchically clustered and overdispersed non-gaussian continuous outcomes for comet assay data. Pharm Stat. 2012; 11: 449-455.
13. Efendi A, Molenberghs G, Njagi EN, Dendale P. A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. Biom J. 2013; 55: 572-588.
14. Dunson DB, Watson M, Taylor JA. Bayesian latent variable models for median regression on multiple outcomes. Biometrics. 2003; 59: 296-304.
15. Dunson DB. Bayesian dynamic modeling of latent trait distributions. Biostatistics. 2006; 7: 551-568.

16. Rodriguez A, Dunson DB, Taylor J. Bayesian hierarchically weighted finite mixture models for samples of distributions. Biostatistics. 2009; 10: 155-171.

17. Hansen MK, Sharma AK, Dybdahl M, Boberg J and Kulahci M. In vivo comet assay statistical analysis and power calculations of mice testicular cells. Mutation Research/Genetic Toxicology and Environnmental Mutagenesis. 2014; 774: 29-40.

18. Collins A, Koppen G, Valdiglesias V, Dusinska M, Kruszewski M, Møller P, et al. The comet assay as a tool for human biomonitoring studies: the ComNet project. Mutat Res Rev Mutat Res. 2014; 759: 27-39.

Cite this article

Verde PE, Rottmann A (2015) Statistical Models to Analyze Genotoxicological Experiments with the Comet Assay Data. Ann Biom Biostat 2(3): 1025.